Malicious Websites Identification Based on Active-Passive Method

Xue-qiang Zou[1,2,3(✉)], Peng Zhang[2], Cai-yun Huang[2,3], and Xiu-guo Bao[1,2,3]

[1] National Computer Network Emergency Response Technical Team/Coordination of China, Beijing 100029, China
[2] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
{zouxueqiang, zhangpeng, huangcaiyun, baoxiuguo}@iie.ac.cn
[3] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract. Nowadays, massive numbers of malicious websites are endeavored to change their hosts/IP addresses to avoid tracking. This paper fills a gap in the study of tracking this kind of websites and offers approaches to detection and identification by combining both active and passive methods. The active method, as bootstrap, is based on crawling traffic from Internet, we can extract title, keywords and picture as features and store them as feature sets. What we do in passive filtering is to match online traffic using the feature sets. Other than finding out those malicious websites, we can extract extra features such as cookie and users information, which is unavailable by using active method, from online traffic and add them to the feature sets created by proceeding active method. According to the experiment, we can have 95.43% true positive rate and 3.90% false positive rate under real data flow in this way.

Keywords: Website identification · Feature extraction · Active detect · Passive monitoring

1 Introduction

With the rapid development of the Internet, websites have become the most important way for people to obtain information. At the same time, many malicious websites spread information such as fraud, violence, pornography, etc. through the Internet, posing a serious threat to cyberspace. Therefore, analyzing and extracting the effective features of websites can identify malicious websites more accurately, reduce the spread of malicious information, and improve the security of network.

At present, some malicious websites are struggling to survive by changing the domain name and IP address. On the one hand, the traffic information of all the domain names under the website cannot be captured by directly matching the known host and IP address. On the other hand, for a website that uses a LAN or a cloud server, its IP address is probably not fixed. For the website service provider, the value of the host

X. Yun et al. (Eds.): CNCERT 2018, CCIS 970, pp. 109–121, 2019.
https://doi.org/10.1007/978-981-13-6621-5_9

field of the HTTP message can be changed by itself, which also makes it possible to forge host information. For the above reasons, the way to directly match the host and IP does not guarantee that the traffic matching the hit is the traffic to the target website.

In view of the above problems, this paper proposes a new method for recording and identifying specific website features. The main research results include:

(1) Based on the identification of website access traffic by matching host and IP attributes, other feature information, including keywords, title, logo picture and cookie, are introduced. Firstly, the cookie is used for website identification. Then, the simhash is used to calculate the value and the similarity comparison is performed by using the simhash algorithm. Finally, the comparison hashing algorithm is used to perform the falsification operation on the comparison result through the logo image of the website. This can eliminate the impact of fake host and IP address uncertainty on website identification, and improve the accuracy of website identification.
(2) According to the characteristics of the website in the traffic, a system of extracting and identifying the active module and passive module linkage is established. After the active discovery of some features, the features are complemented by passive detection, and the host information of the same website whose domain name or IP constantly changing is found. The information is recorded and input into the active acquisition module for further analysis.

The second section of this paper mainly introduces the related work of website detection and identification. The third section describes the extraction system based on active-passive combination, and introduces the way to determine the similarity of the website. The fourth section uses the experiment to quantitatively evaluate the results obtained by the model; Finally, we summarize the paper and point out the next step.

2 Related Work

The current research on automatically identify websites is mainly based on three types of methods: blacklist, active detection and passive monitoring. The blacklist-based technology is mainly implemented by maintaining a blacklist with IP address and domain name. It can be judged the malicious website by checking whether the domain name/IP address appears on the blacklist. Active detection mainly uses crawler to obtain static data, such as keyword in webpage html, and then analyze it through machine learning. Passive detection allows direct analysis of online traffic.

Specifically, some browsers use a built-in blacklist to provide users with lightweight, real-time malicious web page identification services to meet the needs of rapid response. However, most of the blacklists are obtained through manual reporting and client analysis, and the workload is large. Ma et al. [1–3] use machine learning algorithms to identify malicious URLs based on DNS information, WHOIS information, and grammatical features of URLs. Canal et al. [4] added JavaScript and HTML features to improve the accuracy of malicious web page identification by detecting web content. However, such active methods often fail to detect malicious websites in time. In addition, with the popularity of network services, attackers will continue to adopt

some new technical means (for example, automatic domain name generation [5], web page hiding technology [6], etc.) to enhance the concealment of malicious web pages and improve attack efficiency. This further increases the difficulty of malicious website identification.

Therefore, this paper proposes a combination of active and passive methods for malicious website identification. This method can solve the untimely problem of traffic discovery and processing in the active mode, and provides effective features for passive traffic filtering. At the same time, combined with the file content of html and the field content of HTTP head, it can extract more useful information to improve the accuracy of website identification.

3 Proposed Method

There are a large number of malicious websites hiding their existence by constantly changing their domain names and IP addresses. This type of website cannot be detected by filtering traffic directly by specifying the IP or domain name. Here, this paper proposes a method of detecting with the combination of active and passive. By establishing the feature information set of the website, the traffic of particular dynamic website is detected to achieve the purpose of identifying the website. The way of collecting website feature information is mainly divided into two types: active acquisition and passive acquisition. The basic framework of the model is given in Fig. 1. The following sections will illustrate the main parts of the framework.

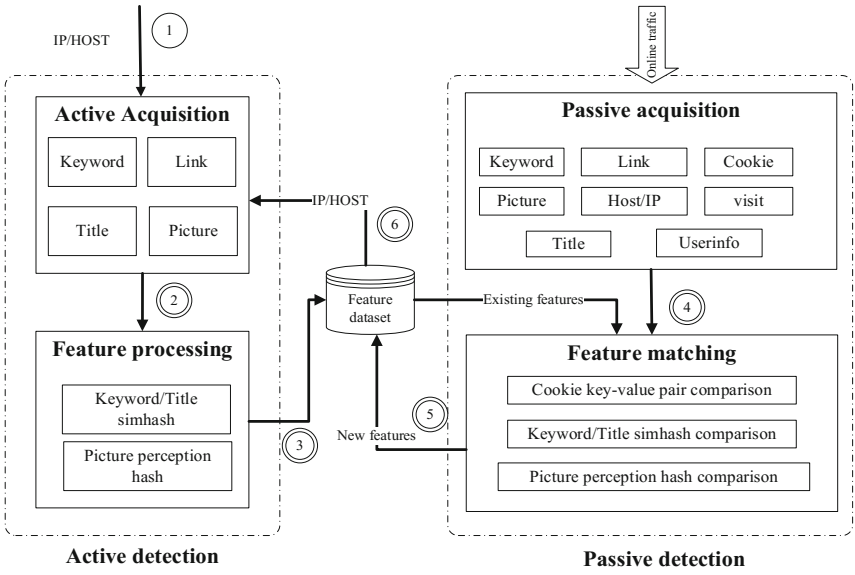

Fig. 1. Website feature feedback identification model

3.1 Active Detection

HTTP response message content plays an important role in network traffic analysis. On the one hand, it truthfully reflects information such as web page content during transmission; on the other hand, HTTP messages have a strict encapsulation format to facilitate data extraction.

Traffic active acquisition is the guide part of the whole model. It mainly sends out GET requests to specific websites by the operator, and extracts the useful information to identify specific websites based on the return HTTP response message content. These feature information mainly include the following two kinds:

(1) HTTP message header field information, such as server, last-modified, x-powered-by and other field information;
(2) HTML file content information, such as title information in HTML, meta information, general software system information, link information, and so on.

These feature information is then saved to a local database to form feature information set for the particular website. At the same time, the crawling data may have some noise, such as the display result is garbled due to different coding methods of the website, and the href tag in the webpage points to the intra-site URL. The active traffic acquisition will pre-process the data before it is saved, such as performing corresponding decoding operations on the webpage content and complementing the link content in the webpage will help to reduce the noise in the dataset.

Specifically, this article selects the following factors that may have an impact on identification accuracy:

Links. For the malicious websites detected, even in the case of different domain names, the number of links in the web page is substantially the same. Although the number of links may vary with the update of the website, the range of changes is relatively limited. Because the number of links on the website shows the page layout of the page from the side, and even if a website is updated, the layout of the website is similar.

Keywords. Keywords generally represent the core content of a web page and are used for retrieval by search engines. Although some sites do not have keywords, there should be similarities for keywords on the same site for sites with keywords.

Title. The title tag in the Html header is used to describe the title information of the website. Usually the same website will have the same title information. Since the title information is not necessary, similar to the keyword attribute, the title information should be used as the identification information for a specific website.

Content Publisher. The author field in the Meta tag is used to convey the publisher information of the content of this page to the search engine. Although the same content publisher may post on different websites, it is also possible that the content publisher has the same name, which makes it impossible to directly identify the website or falsify the website through this option, but this is a property of the website itself. At the same time, the inclusion of content publisher information in the site's feature information set can also provide conditions for similar site queries in the future.

Logo. Most websites have their own logo patterns. Since these logo images may have different height and width, the picture information can be extracted and the two-image similarity calculation is performed using a perceptual hash algorithm.

Host/IP. Recording known URLs for specific websites visited does not reflect the characteristics of such URLs, but helps to more accurately filter specific URL traffic during passive traffic acquisition.

3.2 Passive Detection

Passive traffic detection mainly obtains online traffic information by accessing ISP traffic. Compared with active traffic detection, the feature information obtained through passive mode includes the following points:

(1) The IP address and port, the agent of the sending and receiving parties, and so on, are available in the active acquisition by the operator itself, or are preset by the operator, which is not characteristic.
(2) The information sent to the server when requested by the client, such as cookie information. Since the operator can set the value of the information he wants to send, this information cannot be recorded as part of the signature database in active traffic acquisition. At the same time, this information is also reinforcing for the judgment of the website.
(3) Statistics. The active requester for the service is only the operator's individual, but once the overall traffic is available, the overall request information can be statistically analyzed by sampling. For example, the amount of visits, users accessing the website, and the ratio of traffic to the site.

When the operator obtains a partial information set of a specific website from the active traffic detection, the feature information is pushed to the passive traffic detection. Passive traffic detection compares these feature information with the online traffic it has acquired. If the features can be matched, it is regarded as obtaining the traffic accessing the specific website, and the above three kinds of information can be sequentially recorded and added to the feature information set of the specific website.

Specifically, passive traffic acquisition module mainly records and adds the following feature information:

User. User information including the source IP address of the TCP layer, source port information, and user login time information. Although the user information cannot mark the website, the user-related information can be used to monitor and analyze the IP access traffic in the future to obtain similar sites.

Traffic. These two values are recorded as statistical information in the feature information set. As attribute information that cannot directly filter the traffic filtering, recording it has a certain effect on the network mapping work.

Cookies. Cookies are mainly used by the server to record the basic information of the visitor and some statistics of the access. Chen et al. [7] provide a way to judge the same website through cookies. When there are two or three cookie key pairs in the passively acquired traffic, it can be judged that the traffic belongs to the same website.

Additional Features. Initially, the feature information of active traffic acquisition, such as keywords, title information, etc., may not be complete. The use of passive traffic can obtain more information about the specific feature of a specific website and complement it, so that a higher rate of flow matching can be obtained in the next screening.

Host/IP. The known host and IP information is limited. Using passive traffic for matching may result in more host and IP address information for the homologous website, which can continue feature extraction in the next active acquisition.

3.3 Homologous Website Identification

Homogenous website refers to a website that provides the same content although the domain name or IP address changes. Such websites have similar feature information. For the identification of such websites, it is necessary to use the specific attributes in the feature set to detect and filter online traffic after the feature information set of the website is generated. At this time, features that can be used to judge the similarity of the website include: keyword, title, logo hash value, and cookie.

Based on the above information, the identification process of the homologous website proposed in this paper is shown in Fig. 2.

First, the cookie is selected as a strong feature to identify the website. When there are three or more cookie key-value pairs in the test website traffic, it is determined to be a homogenous website. However, experiments have shown that the cookie value of the same website may be completely different. Therefore, when using cookies can't judge the flow of the passive traffic to the target site, the similarity of the string generated by the keyword and the title, as well as the similarity between the logo pictures of the website can be used. Finally, the hash value of the logo image is used to falsify the homologous websites determined by keywords and titles. The homologous websites with logo pictures are similar in image features. When the keyword and the title information are similar, if the passive traffic does not find a picture similar to the feature set, it can be considered that the detected website is not the homologous site of the target website.

In the following, the identification criteria are described in terms of features, wherein the feature information set includes feature information represented by A, and the corresponding feature information extracted from the traffic is represented by B:

Step1: For cookies, select T_{cookie} as a critical value. When the same number of cookie keys reaches the critical value, it is judged to be the same website [7], which is recorded as $c_{sim}(A, B)$. The method of determining the similarity between two websites by using the cookie key-value for the same number is as shown in the formula (1):

$$c_{sim}(\mathrm{A}, \mathrm{B}) = \begin{cases} 1, & |\mathrm{A} \bigcap \mathrm{B}| \geq T_{cookie} \\ 0, & |\mathrm{A} \bigcap \mathrm{B}| < T_{cookie} \end{cases} \tag{1}$$

Where $|\mathrm{A} \bigcap \mathrm{B}|$ indicates the number of identical key-value pairs of the cookie in the signature database and the cookie detected in the traffic.

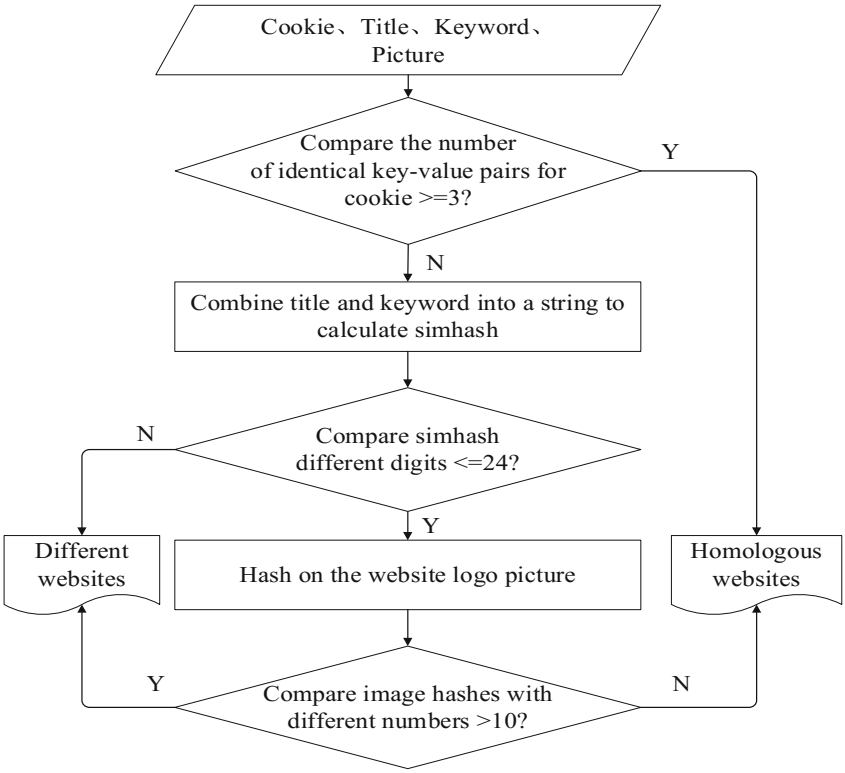

Fig. 2. The process of homologous website identification algorithm

Step2: For keyword and title information, we used LSH (locally sensitive hash) [8] to combine the keyword and title information of the website into a input string. Then extract the keywords that identify the website from this string. These keywords are given a specific weight, and each keyword is hashed to obtain a 64-bit binary string. Multiply the hash value of each keyword by its weight, and add the results of each keyword in bits. Finally, if a value obtained by bit is greater than 0, the bit is marked as 1; otherwise it is marked as 0. For keyword and title information, we used a locally sensitive hash algorithm [8] to combine the keyword and title information of the website into a string input. Then extract the keywords that identify the website from this string. These keywords are given a specific weight, and each extracted keyword is hashed to obtain a 64-bit binary string. Multiply the hash value of each keyword by its weight, and add the results of each keyword in bits. Finally, if a value obtained by a bit is greater than 0, the bit is marked as 1, otherwise it is marked as 0. Then, a 64-bit binary fingerprint based on the website keyword and title is finally generated. When the fingerprint extracted in the passive traffic differs from the fingerprint in the signature database by less than a certain distance, it is determined to be a homogenous website.

Step3: For the logo picture, this paper uses the perceptual hash algorithm [9] to obtain the hash value of the logo image in the active traffic acquisition. When the

passive traffic is acquired, the same operation is performed on the image in the traffic to obtain the hash value, and then compares the hash value of the passively acquired images with the hash value in the feature set to obtain the similarity of the image. In the perceptual hash algorithm, the image is reduced to an 8 * 8 size, so it is converted into a 64-level grayscale image, and the grayscale average of the 64 pixels is calculated. The gray value of each pixel is compared with the average value and then 0 and 1 are respectively recorded. This gives a 64-bit number that is the fingerprint of this picture. The similarity index of the two picture hashes, that is, the different data bits of the hash value (Hamming distance) is shown in formula (2):

$$\mathrm{pHash}\left(\mathrm{A}_{pic}, \mathrm{B}_{pici}\right) \quad i\epsilon[1, n] \tag{2}$$

The hash value of the image stored in the feature set is compared with the hash value of the image contained in the online traffic, where n in the formula (2) represents the number of all the pictures on the online traffic. The minimum value of the two differences is chosen as the basis for judging the similarity of the two websites, which is denoted as $\mathrm{p}_{sim}(\mathrm{A}, \mathrm{B})$. Select T_{pic} as the critical value, and use the Hamming distance to determine the similarity between the two websites as shown in formula (3):

$$\mathrm{p}_{sim}(\mathrm{A}, \mathrm{B}) = \begin{cases} 1, & \left|\mathrm{pHash}\left(\mathrm{A}_{pic}, \mathrm{B}_{pici}\right)\right| \leq T_{\mathrm{pic}} \\ 0, & \left|\mathrm{pHash}\left(\mathrm{A}_{pic}, \mathrm{B}_{pici}\right)\right| > T_{\mathrm{pic}} \end{cases} \tag{3}$$

4 Experimental Evaluation

4.1 Data Collection

In order to evaluate the effectiveness of the active-passive combination detection method, this paper conducts deployment experiments according to the model diagram shown in Fig. 1, and performs data collection in a real Internet environment. In terms of active detection, this article has conducted several random interviews on websites such as the Caoliu Forum. Use the script program written in python 2.7.11 to get the corresponding keyword value on the website and save it to the database after processing. For passive detection, this paper obtains passive data acquisition and analysis by accessing certain ISP traffic.

The data actively obtained in this paper mainly comes from three adult websites whose names/IP addresses will change, such as the Cailiu Forum, Maya Forum, and sis001, and the data is mainly crawled for the website homepage. Compared with the method of detecting the website by using only the keyword/title information, the method of this paper can well realize the identification of the website without the keyword/title; and the way of using only the domain name/IP address Compared with the method of detecting the website, the method of this paper can well realize the identification of the website with the domain name/IP change.

4.2 Parameter Setting

The models used in this paper include the active detection phase and the passive filtering rediscovery phase. The most important measurement metrics are the accuracy of the final passive filtering website and the false positive rate of the website, as described in 4.3. At the same time, as rediscovery progresses, the data in the feature set will become more and more, and the results obtained will be more accurate.

In the experiment, the value of the same cookie key-value to the number of the number, the value of the hash value of the hash value of the logo picture and the value of the simhash value of the keyword and the title information synthesis string will have an effect on the accuracy of the final judgment of the specific website.

Since the cookie information is used as a strong feature to judge the website, so when the number of the same key-value pairs is greater than or equal to 3, the site of the passive traffic is judged to be the homologous website of the target website [7]. At the same time, when the similarity between two pictures is calculated by the perceptual hash algorithm [9], when the hash value of the two pictures is less than 5, it is judged that the two pictures are basically the same picture; When the Hamming distance of the hash value is greater than 10, it is judged that the two pictures are basically different. Here, because the picture information is used for falsification, in order to ensure a certain recall rate, the picture hash Hamming distance 10 is selected as the demarcation value.

4.3 Evaluation Method

True Positive Rate (TPR), False Positive Rate (FPR), Accuracy and Recall are the common evaluation indicators for classification effects, considering the characteristics of malicious website identification in real traffic environment. We choose TPR and FPR that is ROC curve, as the evaluation criteria.

For a category of Class, the category that belongs to the category is usually called positive, and the one that does not belong to the category is usually called negative, as shown in Table 1.

Table 1. Double classifier model

	True	False
Positive	TP	FN
Negative	FP	TN

For a certain class, those belonging to this category are usually called positive cases, which are not usually called negative cases, as shown in Table 1.

Where TP indicates the number of identified in the positive case, and FN indicates the number identified in the negative case, and FP indicates the number that is not

identified in the positive case, and TN indicates the number that is not identified in the negative case. The definition of TPR and FPR is shown in formula (4) and formula (5):

$$\text{TPR} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{5}$$

At the same time, the area AUC (Area Under Curve) under the ROC curve is used to judge the effect of the classifier. The larger the AUC, the better the classifier.

4.4 Experimental Result

Among them, the blue line indicates the ROC curve obtained by using the homologous website identification algorithm proposed in this paper, and the orange line indicates the ROC curve obtained by direct matching with the host. It can be intuitively seen that the AUC obtained by the simhash algorithm is much larger than the AUC obtained by using the host, which proves that the proposed algorithm has a better effect.

For the selection of the hash value of the final keyword and the title synthesis string, we tested the different URLs of the Cailiu Forum, Maya Forum and sis001 websites, and obtained the values under different simhash (A, B). The TPR and FPR identified by the website are shown in Fig. 3.

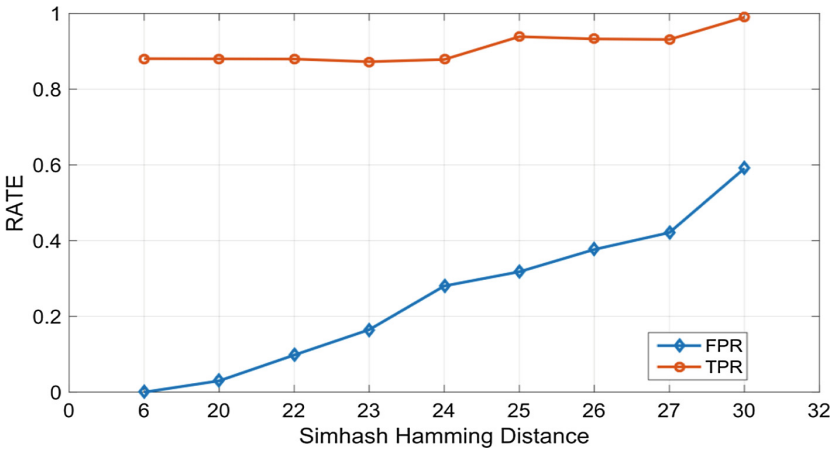

Fig. 3. TPR and FPR graphs obtained using simhash (Color figure online)

Among the Fig. 3, the blue line indicates the FPR obtained by simhash, and the orange line indicates the TPR obtained by the same way. It can be seen that the greater the value of the simhash value, the higher the TPR for the identification. But at the same time, the FPR also has a big rise in the trend. When the cutoff value of the

simhash value is 30, the TPR can reach 99%, and the FPR will reach 59.09%. However the Logo will be used for falsification later, when the simhash cutoff value is 24, the TPR obtained is 93.93% and the FPR is 31.81%, the growth rate of FPR is less than the growth rate of TPR.

After adding the cookie and the logo information to judge, the method is compared with the method of identifying the website through host, and the obtained result is shown in Fig. 4.

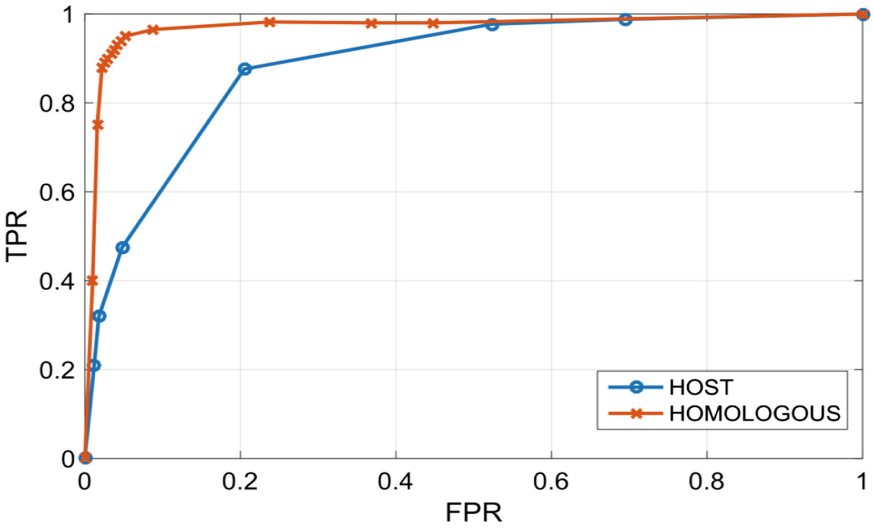

Fig. 4. Identification comparison of homologous website algorithm with host (Color figure online)

In Fig. 4, the blue line indicates the ROC curve obtained by using the homologous website identification algorithm proposed in this paper, and the orange line indicates the ROC curve obtained by direct matching with the host. It can be seen intuitively that the AUC obtained by the simhash algorithm is much larger than the AUC obtained by using the host, which proves that the proposed algorithm has a better effect.

5 Conclusion

In view of the identification of specific websites, this paper uses the combination of active and passive to discover and record the key attributes of specific websites based on the content of http header and html files, and proves that this method is validity and accuracy to identify the website especially when the domain name/IP changes.

We have found that user detection and content publisher information for accessing specific websites can be reversed detected. For visitors, it is possible to capture the source IP/port as the traffic of the visitor and perform keyword matching on the access traffic, because the visitor often knows not only a specific website but also such a

website. Have some understanding. For content publishers, when they post a malicious website, we have reason to believe that he has a tendency to post other malicious websites. Therefore, in the next step of the work, the two will be monitored to try to get the same type of malicious website.

Not only the user information, but also the statistical information obtained by analyzing the feature data set of a specific website has a great effect. For example, statistics on websites with domain name/IP changes are more likely to use information such as what kind of framework/server to build, and the threshold for the number of links to such sites. Although this information cannot be directly used to identify a website, it has a certain contribution to the identification of a website because it has these attributes. Later, through the machine learning method, the website feature information set can be directly reduced in dimension, and the identification with higher accuracy can be obtained. At the same time, statistics on the number of users and the number of visits to such sites can be used to estimate how much impact such sites can have.

Acknowledgment. The research work is supported by National Key R&D Program of China (No. 2017YFB0801701) and National Natural Science Foundation under Grant (No. 61300206).

References

1. Ma, J., Saul, L.K., Savage, S., Voelker, G.M.: Beyond blacklists: learning to detect malicious web sites from suspicious URLs. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, pp. 1245–1253 (2009)
2. Ma, J., Saul, L.K., Savage, S., Voelker, G.M.: Identifying suspicious URLs: an application of large-scale online learning. In: Proceedings of the 26th Annual International Conference on Machine Learning (ICML), Montreal, Quebec, Canada, pp. 681–688 (2009)
3. Ma, J., Saul, L.K., Savage, S., Voelker, G.M.: Learning to detect malicious URLs. ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 1–24 (2011)
4. Canali, D., et al.: Prophiler: a fast filter for the large-scale detection of malicious web pages. In: Proceedings of the 20th International Conference on World Wide Web (WWW), Hyderabad, India, pp. 197–206 (2011)
5. Yadav, S., Reddy, A.K.K., Reddy, A.L., et al.: Detecting algorithmically generated malicious domain names. In: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement (IMC), New York, NY, USA, pp. 48–61 (2010)
6. Kolbitsch, C., Livshits, B., Zorn, B., et al.: Rozzle: de-cloaking internet malware. In: Proceedings of the IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, pp. 443–457 (2012)
7. Chen, Z., Zhang, P., Zheng, C., Liu, Q.: CookieMine: towards real-time reconstruction of web-downloading chains from network traces. In: 2016 IEEE International Conference on Communications (ICC), pp. 1–6. IEEE (2016)
8. Manku, G.S., Jain, A., Das Sarma, A.: Detecting near-duplicates for web crawling. In: International World Wide Web Conference (2007)
9. Schneider, M., Chang, S.-F.: A robust content based digital signature for image authentication. In: Proceedings of IEEE International Conference on Image Processing, Lausanne, Switzerland, vol. 3, no. 3, pp. 227–230 (1996)